Philos Stud (2013) 162:421–445
DOI 10.1007/s11098-011-9775-5

The armchair and the trolley: an argument for experimental ethics

Guy Kahane

Published online: 11 August 2011

Abstract Ethical theory often starts with our intuitions about particular cases and tries to uncover the principles that are implicit in them; work on the 'trolley problem' is a paradigmatic example of this approach. But ethicists are no longer the only ones chasing trolleys. In recent years, psychologists and neuroscientists have also turned to study our moral intuitions and what underlies them. The relation between these two inquiries, which investigate similar examples and intuitions, and sometimes produce parallel results, is puzzling. Does it matter to ethics whether its armchair conclusions match the psychologists' findings? I argue that reflection on this question exposes psychological presuppositions implicit in armchair ethical theorising. When these presuppositions are made explicit, it becomes clear that empirical evidence can (and should) play a positive role in ethical theorising. Unlike recent assaults on the armchair, the argument I develop is not driven by a naturalist agenda, or meant to cast doubt on the reliability of our moral intuitions; on the contrary, it is even compatible with non-naturalism, and takes the reliability of intuition as its premise. The argument is rather that *if* our moral intuitions are reliable, then psychological evidence should play a surprisingly significant role in the justification of moral principles.

Keywords Moral intuitions · Empirical psychology · Ethical theory · Experimental ethics · The trolley problem · Moral principles · Moral epistemology · Armchair knowledge

1 Introduction

If you saw a gang of naturalists pouring gasoline over an armchair and setting it on fire, would you think it wrong? The argument I will develop condones no such

G. Kahane (✉)
Oxford Uehiro Centre for Practical Ethics, Littlegate House, St Ebbe's St, Oxford OX1 1PT, UK
e-mail: guy.kahane@philosophy.ox.ac.uk

behaviour. It is not driven by any naturalist agenda—it is meant to be compatible even with the most robust forms of non-naturalist realism. And, unlike recent assaults on armchair theorising, this argument casts no doubt on our intuitions; on the contrary, it sets out from the assumption that our intuitions are largely trustworthy, and asks what follows. Yet from this widely held assumption it leads, in several simple steps, to conclusions that many will find surprising, even disturbing. If these conclusions are correct, then we may need to do ethics differently. There will be no auto-da-fé, but empirical research will take over part of ethics. The armchair will be pushed to the corner. This is not an outcome I especially welcome, but I do not see how it can be avoided.

1.1 Puzzling parallels

Normative ethics aims to provide principled answers to normative questions about what we ought to do. It often starts with intuitions about particular cases and attempts to uncover the general moral principles that underlie these intuitions.[1] For example, in trying to uncover deontological constraints on permissible killing, ethicists have notoriously considered numerous variations on the Sidetrack case, where a bystander can save five innocents from being crashed to death by a runaway trolley by switching it to another track, where it will kill only one. Whereas it seems to most that saving the five here is permissible, far fewer think that we are permitted to kill one to save five in the Footbridge case, where one needs to push a large person off a footbridge and onto the trolley's track. Ethicists seek to discover what principle might underlie this intuitive distinction.[2]

But ethicists are no longer the only ones playing with trolleys. They have recently been joined by psychologists and neuroscientists who have begun to use behavioural experiments and neuroimaging to study the intuitive responses of non-philosophers to these very same trolley cases, aiming to identify the psychological mechanisms and neural processes that generate these intuitions.[3]

One influential principled explanation for the intuitive distinction between Sidetrack and Footbridge is the *doctrine of double effect* (DDE), on which it is impermissible to intentionally cause harm, but permissible to cause harm as a foreseen but unintended consequence of one's action. The DDE has been widely criticised, but variants of it are still vigorously defended. The DDE, however, has also attracted support from within psychology: psychologists have recently claimed, on the basis of empirical evidence, that the DDE indeed underlies our intuitions about the core Trolley cases.[4]

[1] With others, I'll use 'intuition' to refer to propositional seemings (cf. Huemer 2006, 102) or dispositions to non-inferential belief. I'll later consider the significance of restricting the relevant class of intuitions to 'considered' ones. And unless stated otherwise, I'll take intuitions to be about *particular* cases.

[2] Thomson (1985), Kamm (2006).

[3] Cf. Greene et al. (2001, 2004), Cushman et al. (2006) and very many more.

[4] Cushman et al. (2006).

So ethicists and scientists are now investigating the very same trolley problems and the intuitions that they elicit, and sometimes even come up with what can appear to be identical explanations. But there is also plenty of divergence. Ethicists have discussed numerous trolley variants, and candidate principles that might explain them, which (as at least as of now) no psychologist has considered (Frances Kamm has even discussed a Doctrine of Triple Effect). And on the other hand the neuroscientist Joshua Greene has proposed that what explains our different responses to Sidetrack and Footbridge is an emotional revulsion to 'up close and personal' violence to others, a view few if any ethicists would take to be a serious candidate for a vindicating explanation of the distinction.[5]

At least some moral philosophers are happy to think about moral intuitions in terms that draw on psychology. Kamm, for example, suggests that

> [our moral] responses come from and reveal some underlying psychologically real structure, a structure that was always (unconsciously) part of the thought processes of some people. Such people embody the reasoning and principles (which may be thought of as an internal program) that generate these responses. The point is to make the reasons and principles explicit.[6]

But this apparent overlap in the object of study and occasional results of ethical theory and psychology is puzzling, and Kamm's remarks do little to clarify the relation between ethical theory and the psychology of morality. Should we expect, as these remarks suggest, that as these parallel inquiries continue, their results would strongly converge? But if so, then in what sense are they really distinct inquiries? And if they are distinct inquiries, would it really matter whether or not there is such a convergence? Suppose that psychological inquiry reveals that the content of the underlying 'internal program' bears *no* resemblance to the moral principles identified, from the armchair, by Kamm or others. Would this make no difference to ethical theory?

1.2 Is ethical theory just a branch of empirical psychology?

One understanding of the analogy Rawls has long ago drawn between ethical theory and linguistics suggests a simple answer.[7] There would be nothing puzzling, and a simple explanation of convergence, if the relation between ethical theory and empirical moral psychology was one of simple *identity*. Thus Susan Dwyer, one of the proponents of this 'linguistic analogy', writes that

> ethical theory is concerned with moral competence, with giving an account of what makes moral judgement and behaviour possible. Individuals' moral judgements—their moral intuitions—are data for moral theory in the same way that acceptability judgements are data for linguistic theory, in the sense

[5] Greene et al. (2009) now offers a somewhat different explanation.

[6] Kamm (1993, 6).

[7] Rawls (1971).

that moral intuitions about a particular case can lend support, or detract from, a given moral principle.[8]

The idea seems to be that ethics is simply the more abstract part of moral psychology—that it's concerned with describing the abstract computations that are implemented by some neurocognitive system.[9] But the aim of ethical theory is surely not to investigate moral competence, or people's psychology or capacities, but to answer substantive normative questions. Moral intuitions may be *evidence* for or against possible answers to these questions, but they are not *data* that moral theories seek to causally *explain*. The aim of ethical theory is to get things right, not to explain why we have a certain set of beliefs (let alone of intuitions). Moral theory is a normative enterprise, empirical moral psychology a descriptive one. The first looks outward, seeking correct answers to normative questions; the other looks inwards, into a person's psychology.[10]

2 Intuitions and principles in ethical theory and in psychology

The differences between ethics and psychology emerge even more sharply once we go beyond surface similarities and spell out the distinct senses in which these two inquiries seek to uncover the general principles that underlie our moral intuitions.

2.1 Intuitions in ethical theory

Consider again the example of the DDE. In one formulation, it is the claim that

> *Doctrine of Double Effect*. It is morally impermissible to intend to harm an innocent person, but it is morally permissible to act out of an intention to bring about good consequences, when harm is a foreseen but unintended consequence of that act.

The DDE makes a normative, not descriptive claim. In essence, the DDE claims that there is an intrinsic moral difference between intending and merely foreseeing harm. Ethicists aim to find out whether moral principles such as the DDE are *true* or *correct*, and to integrate them, if possible, in more comprehensive theories.[11]

The DDE makes no mention of intuitions, and it is not *about* anyone's intuitions. This is not to say that there is no relation between the DDE and our intuitions. There

[8] Dwyer (1999, 180).

[9] Mikhail (2007) makes a similar suggestion, but refers not to linguistics but to David Marr's influential framework for cognitive science.

[10] I emphasize the distinction between normative and descriptive aims, but the more fundamental distinction is actually that between seeking the *truth* about some domain and seeking to understand the psychology that underlies our thinking about a domain—a distinction that also applies in non-normative domains.

[11] Ethicists seek true moral principles, but this needn't commit them—or my argument—to any kind of moral realism; for further discussion, see Sect. 5.3 below.

is a sense in which the DDE can be said to *explain* our intuitions. Those who accept the DDE also usually accept that

> The DDE *grounds*, and thus *explains*, the moral difference between Trolley and Footbridge.

The DDE might similarly explain the truth of a range of intuitively compelling moral propositions about killing and harm. These more specific propositions can be *derived* (whether deductively or in some looser fashion) from the DDE, a general moral principle.[12]

Notice that what the DDE might explain is the truth of the *propositional content* of some of our intuitions. It does not (and could not) explain why we *have* intuitions with these contents.[13]

The DDE, if true, might explain the propositional content of some of our intuitions—if *these* are true. But, conversely, these intuitions can also give us epistemic *reasons* to *believe* in the DDE; they might be *evidence* for *its* truth[14]:

> *Local Coherence*. We have defeasible reason to believe in the DDE because, out of relevant possible principles, it is the one that best coheres with the propositional content of our intuitions about permissible harm and killing.

Coherence is in the first instance a logical relation between propositions. However, what supports *belief* in the DDE is the fact that the relevant set of propositions is one we also *believe*, on intuitive grounds. Thus, although some principle might cohere with some set of moral propositions about particular cases, this in itself gives this principle no epistemic support. Indeed, a principle wouldn't have such support even if it perfectly cohered with our intuitions, if these are intuitions we have reason to reject.

That local coherence with our moral intuitions gives such epistemic reasons is very widely accepted.[15] These reasons are defeasible. They will be defeated if the evidential status of the relevant intuitions has been defeated. And the principle that best coheres with a set of intuitions might also be one that is intrinsically implausible, or coheres badly with other established principles or moral values, or would lead to a general moral theory that is less simple than competing alternatives.

[12] The DDE might be a fundamental, underivable moral truth. Or if its truth can itself be explained, this would be by appeal to an even more general principle, such as the Kantian idea that we ought to treat persons as ends and not merely as means.

[13] The DDE might also, in one sense, explain why we have these intuitions if (i) the DDE is true and (ii) our intuitions track the truth. But an intuition might be true, and its truth explained by the DDE, yet also be accidental.

[14] If we come to believe in the DDE as a result of reflection on our intuitions, then, in one sense, our intuitions could also be said to *explain* why we *hold* the DDE to be true. Here 'explain' is used in a sense that is both causal and rational: our treating out intuitions as evidence is a rationalizing explanation of why we come to believe the DDE.

[15] What I call 'local coherence' is obviously similar to the early stages of narrow reflective equilibrium (though see fn. 29 for a stronger claim). But I'll avoid using this term because I aim to describe a commonly accepted practice, not some theoretical view that may or may not accurately reflect this practice, and which might be associated with contentious metaethical and epistemic commitments.

Any of these 'global' considerations could defeat the intuitional justification of a principle.[16]

So our intuitions can justify belief in a principle, which in turn grounds and thus explains these very intuitions. But things don't only go in such a circle. To start with, not all intuitions survive reflection—we can justifiably dismiss those intuitions that don't cohere with the principle that best coheres with our intuitions. (As is familiar, this process can go through several iterations until equilibrium is achieved.) And more critically, the principle we arrive at will typically also imply conclusions about particular cases where we lack strong (or any) intuitions. Here, the general principle both explains moral propositions with more particular content *and* is the source of our reasons to believe that they are true. (Indeed, if we do come to believe in these specific implications, our belief in the principle would also *causally* explain why we thus believe.)

2.2 Intuitions in psychology

The relation between intuitions and principles takes a rather different form in the psychology of morality. When psychologists set out to explain what underlies the distinction between Sidetrack and Footbridge, they do not seek to explain why such a distinction holds, or whether it does. What they seek to explain is why we (or some of us) draw this distinction, or, more accurately, why we have the intuitions that draw us to draw this distinction. Psychologists are interested in *causal* explanations of *psychological states*, and not necessarily in explanations that rationalize those states.

In the first instance, psychologists seek to establish causal regularities of a functional kind: causal claims that relate patterns of non-moral input to patterns of moral output. Thus, by comparing the differential responses of participants in experiments to a range of moral scenarios, psychologists might conclude, for example, that

> *Underlying Causality*. The fact that Footbridge (and similar cases) involves intending to harm and Sidetrack (and similar cases) involves merely foreseeing to harm causally explains why these scenarios generate opposing intuitions about moral permissibility.

This causal explanation cites differences in the content of moral scenarios but of course this causal regularity is mediated by the fact that these differences in the content of the moral scenarios are *perceived* or *cognitively registered*. Our cognitive system essentially performs a mapping function that takes as input certain non-moral properties (relating to intentions, acts, consequences, etc.) and generates as output certain moral intuitions: it might map, for example, acts that involve the intention to harm onto a verdict of wrongness, and acts that involve merely foreseeing harm onto a verdict of permissibility. When this is true, we can say that

[16] Such global reflection might lead us to reject our intuition that pushing in Footbridge is wrong, as utilitarians do, or even to reject the intuition that switching in Sidetrack is permitted (cf. Thomson 2008).

the DDE is being *tracked* by our intuitions, or is *causally operative* in generating them.[17]

So there is also a psychological sense in which the DDE could be claimed to explain our pattern of intuitions, and in which the evidence supporting this explanation would be provided by our pattern of intuitions.[18] In *Underlying Causality*, the properties cited by the DDE, *as* cognitively registered by persons, causally explain why these persons have some pattern of moral intuitions. This is very different from the claim that the DDE explains why certain propositions are true; the psychological claim might be true even if the DDE is itself false.

The identification of the principle that best coheres with some set of intuitions is only one stage in ethical reflection. Ethicists are ultimately interested not in local coherence but in truth, and in finding, if possible, general theories that would unify the different principles we've singled out through reflection on our intuitions. Similarly, the identification of functional causal generalizations is merely the first stage of psychological inquiry. Psychologists want not just to identify such generalizations, but also to explain *why* and *how* they hold. Such explanations would in part involve uncovering the subpersonal causal mechanisms that realize these mapping functions:

> *Subpersonal Realization*. Mechanism such-and-such explains why we respond differently to foreseeing and intending, and thus our pattern of intuitions about Sidetrack and Footbridge.

Now this kind of explanation can be offered at different levels: as an information processing explanation, setting out the representations and computations that underlie the mapping function, or at the 'hardware' level of its neural realization. And current scientific theorizing about moral psychology offers very different competing accounts of how such functions are realized: this might involve simple emotional responses arising from 'primitive' parts of the brain or elaborate computations in a dedicated 'moral organ'. Importantly, however, we should distinguish disagreement about *Underlying Causality* from disagreement about *Subpersonal Realization*. Greene, for example, has at one point claimed both that what underlies the difference between Sidetrack and Footbridge is the fact that in Footbridge we harm someone in an up-close and personal way *and* that this

[17] (1) Which shouldn't, of course, be confused with the claims that the DDE is true or has causal powers! (2) By 'non-moral input' I mean only the non-moral *factual content* of the relevant scenario. Psychologists are also interested in factors that causally affect moral judgment when this factual content is kept *fixed*—e.g. the effects of style of presentation. I ignore such factors here since they trivially violate supervenience and could not be morally relevant. (3) As the 'Knobe effect' suggests, moral evaluations might have top-down influences on the way the non-moral facts are categorized. But the possibility of such 'loops' is compatible with there being such mapping functions. And there may of course also be cases where moral input is mapped onto further moral output.

[18] *Worry*: Psychologists don't always distinguish moral intuitions (qua propositional seemings/dispositions to believe) from the judgments immediately formed on their basis—and isn't it only the latter that they directly study? *Reply*: Psychological explanations of common non-inferential moral judgments are presumably also explanations of the intuitions that underlie them. Moreover, intuitions in this sense *are* central to psychological theorizing; for example, there is behavioural and neural evidence that even individuals who judge that it's permissible to push the fat man in Footbridge must first overcome an immediate contrary inclination (cf. Greene et al. 2004).

difference is based in a primitive emotional response. By contrast, the psychologist Marc Hauser has argued that something like the DDE underlies this distinction, *and* that our intuitions are generated by complex computations that do not involve emotion. But in principle the DDE might be implemented by some emotional response, and sensitivity to up-close violence by 'cold' (though admittedly simple) cognitive processing.[19] Note also that although the mapping function might be internally represented as an explicit rule—perhaps even encoded in the language of thought—a principle might be causally operative in the above sense even if it's not internally represented.[20]

3 The argument

When psychological research and ethical theory are viewed from up close, it's easy to conclude that they are utterly distinct inquiries. But this impression is mistaken.

My basic argument is simple. When we ask ourselves whether we should switch a trolley to save five, we take our intuition about this possible act to give good reasons for moral belief. We take it that

(1) Our moral intuitions about particular cases give us defeasible reason to believe in their contents.[21]

But ethical reflection doesn't stop here. We also seek general moral principles that would explain our pattern of intuitions about particular cases. We assume that, when not biased, these intuitions systematically respond to the presence and absence of morally relevant properties. In other words,

(2) Our moral intuitions about particular cases *track* certain moral principles.

Our intuitions may tell us that it's permissible to divert the trolley in Sidetrack, but wrong to push the stranger in Footbridge. These acts, however, are surely permissible or wrong in virtue of something, and about this our intuitions are typically silent. But if there are general moral principles, and our intuitions are broadly reliable, then our intuitions should be systematically responsive to the properties cited by these principles.[22] And this means that by identifying the properties to which our intuitions systematically respond, we can identify the principles that are implicit in them. Therefore, so long as we are entitled to assume (1),

[19] Psychologists are also interested in *distal explanations* of patterns of intuitions. Such patterns might hold because they are innate, or through social conditioning, or an internal developmental process. Our answers to *Underlying Causality* and *Subpersonal Realization* can leave this further question open.

[20] Cf. Nichols (2005) for how the DDE might be realized without such internal representation.

[21] Some philosophers understand 'intuitions' to be non-inferential beliefs rather than propositional seemings. But for our purposes, this is just a terminological difference, given that these philosophers still hold that we can justifiedly form such beliefs because their content *seems* non-inferentially credible.

[22] I will refer interchangeably to the properties our intuitions track and the principles that cite these properties.

(3) Evidence about what moral principles our intuitions track gives us defeasible reason to believe in these moral principles. [*From 1, 2*]

But

(4) Facts about what principles our intuitions track are *empirical* facts, and are therefore discoverable using the methods of empirical psychology.

More specifically, facts about what our intuitions track are *exactly* the kind of *functional* facts psychologists seek to discover.

Therefore

(5) Psychological evidence about the principles our intuitions track gives us defeasible reasons to endorse these moral principles. [*From 3, 4*][23]

The argument is meant to apply to anyone who allows that moral intuitions about particular cases can give at least some epistemic support to general principles—to all who accept (1) and (2).[24] Versions of these premises are, I believe, accepted by many ethicists. I'll later briefly discuss approaches that reject (1), and only extreme particularists accept (1) without accepting (2).[25] Notice that doubts about the *extent* to which our intuitions are reliable—doubts perhaps fuelled by empirical evidence about disagreement and diversity—only affect the *scope* of the argument. My argument is simply that *whenever* we are justified in giving weight to local armchair coherence, we should give weight to psychological evidence.

Premise (3) is the key move. It might be wondered how it squares with my earlier description of the role of coherence in ethical theorising. Indeed, it might be argued as follows: 'Relations of coherence between intuitions and principles are logical relations, hence a priori, hence to be discovered in the armchair. If it is coherence with intuitions that supports principles, then no opening is left for empirical input to the process. It could not matter whether the principles ethicists identify converge with those identified by the psychologists. The armchair is safe.'

This objection misrepresents our interest in coherence between intuitions and principles. We accept (3) because we accept (2). That is,

[23] It might be argued that what gives us reasons to believe in these principles are our *intuitions*; evidence about what our intuitions track rather gives us reasons to believe we *have* these reasons, reasons we had anyway. Thus evidence about what someone else's intuitions are tracking might give us a reason to believe that this person has reason to believe the corresponding principle without giving *us* any reason to believe that principle. Since this point would presumably also apply to the armchair search for local coherence, and since it doesn't affect the main thrust of the argument, I will set it aside for simplicity's sake. I am grateful here to an anonymous referee.

[24] I believe that my argument can be made compatible even with radical coherentist views that reject any kind of non-inferential justification, so long as they allow that particular beliefs can justify general principles. But I'll ignore such views since few (if any) ethicists hold them.

[25] More moderate particularist admit that there are more general moral patterns, though they deny these amount to genuine exceptionless moral principles (cf. Dancy 2006). For an empirically-based anti-particularist argument that our moral intuitions must be responding to general patterns, see Jackson et al. (2000).

> We have reason to believe in the principle that best coheres with our intuitions *because* such local coherence is evidence that this is the principle that our intuitions track.

This understanding of our interest in coherence might seem surprising. It is therefore worth considering what it would mean to deny it.

One way to deny it is to think of our aim in ethical theorising as not to uncover pre-existing principles, but to *construct* general principles that would preserve as many of our particular intuitions as possible—this view sees moral principles as no more than ways of organizing or systematizing our particular intuitions, in the way scientific theories are claimed to organize experience on some instrumentalist views.[26]

But such a view coheres poorly with the way most ethicists see their aims, and with actual ethical practice. When ethicists find our pattern of intuitions about trolley cases puzzling, when they wonder what principle might be implicit in it, they clearly don't take such an instrumental attitude to moral principles. Nor do ethicists see systematic regularities in our moral intuitions as merely a fortunate accident which helps us better integrate our diverse responses to particular cases.

Most ethicists do not simply seek the principles that would preserve as many of their existing intuitions as possible. On the contrary, they actively generate intuitions about new cases, cases that might decide between competing principles that seem to cohere equally with the existing set of intuitions. This would make little sense if their aim was mere local coherence. Indeed, one possible consequence of extending the range of intuitions in the set is that we would come to *reject* firm intuitions from the initial set, intuitions that wouldn't best cohere with the principle that best coheres with the larger set. Again, this would make no sense if the aim of ethical theorising was merely to preserve as many of our existing intuitions as possible. Ethicists aren't satisfied with just *any* principle that happens to cohere with our existing set of intuitions—they worry that this pattern is merely *accidental*, that a *different* pattern will emerge when they extend the set.

Then there is the point that ethicists don't admit just any intuition to the set. Intuitions must first go through epistemic screening: they must, for example, be the result of clear and calm reflection, not agitation or self-interested bias. Again it seems that ethicists aren't just interested in preserving our existing intuitions, but in removing noise and interference that might make us perceive in them what is merely a random pattern. In ethical theorising, then, we also want to identify non-accidental patterns in *non-accidental* intuitions.

It might be objected that this screening is pragmatic, aiming merely at stability. We simply want to rule out transient intuitions that we will no longer have later on, or that we are less likely to share with others. But this is implausible. To see this, imagine that a mischievous demon decided to play with our intuitions. He flips a coin and on this basis makes us all respond with certain intuitions when we calmly reflect on certain examples. When these devilish intuitions are joined to our other intuitions, they accidentally best cohere with some principle—say, the DDE—

[26] See Dworkin (1973), for such a 'constructive view'.

which would not be otherwise favoured by our intuitions. It seems to me clear that if we discovered the random source of these common and highly stable intuitions, this intuitive support for the DDE would be defeated. Local coherence isn't sufficient for justification when it's known to be merely accidental.

So what we seek in ethical theorising are non-accidental patterns in non-accidental intuitions. The armchair search for local coherence is a way to detect such patterns, a way to identify the properties that our intuitions are systematically responding to, and to identify and dismiss stray or biased intuitions that obscure this underlying pattern. But this is just is to accept (3) and the understanding of the coherence that I suggested above.[27]

Consider next a different objection. 'The problem with (3) is rather that it assumes a far too simple epistemic step from intuitions to principles. When we engage in ethical theorising, we constantly move back and forth between particular intuitions and principles, as well as between these and other principles and background theories. It is this holistic character of moral theorising that your argument distorts or overlooks, and which blocks the step to (5).'[28]

Premise (3) assumes that facts about the properties that our moral intuitions track give epistemic grounds for believing in the corresponding moral principles: that if our intuitions about trolley cases respond to the distinction between intending and foreseeing, then this gives us reason to endorse the DDE.

When we go back and forth between intuitions and principles, one thing we are trying to do is distinguish those intuitions that are reliable from those who are merely random or biased—and thus avoid endorsing principles based on such irrelevant intuitions. Since the aim here is precisely to single out the properties our intuitions are genuinely tracking, this feature of our practice is hardly an objection to (3). It supports it.[29]

So the issue must really be about epistemic considerations that go beyond local coherence. But my argument is perfectly compatible with accepting other epistemic factors into the mix, including top-down pressures from other principles and values, and from background theories. As I emphasized, the support given by particular intuitions is defeasible. The principle they identify might turn out to be intrinsically implausible or in tension with strongly supported values. And background theories and other considerations (including empirical evidence about bias and diversity) might give us reason to think that some set of intuitions is unreliable or prejudiced, defeating premise (1) for that particular domain. None of this is in tension with my argument.

The argument is also compatible with more subtle top-down pressures. Suppose some set of intuitions track property A, but considerations of global coherence

[27] This understanding of the search for coherence seems to me compatible with Daniels (1979). For the view that coherence is epistemically valuable because it bolsters reliability, see Sosa (2009).

[28] I owe this objection to Simon Rippon.

[29] If this iterative process actually aims to rule out stray intuitions, then what I mean by 'local coherence'—and thus the immediate target of my argument—*wouldn't* be restricted only to the *first* stage of narrow reflective equilibrium (Daniels 1979). Notice also that something like this back and forth would still be preserved in psychological research that aimed to identify the factor being tracked, where hypotheses would be repeatedly generated and tested against intuitions.

support a somewhat different property A*. Depending on the respective epistemic weight we give to local intuitions and to more global considerations, we might have overall reason to accept factor A* over A, and to revise our ground-level judgements accordingly.[30] This doesn't show that there was *no* reason to adopt A, not even that the intuitional support for A didn't play an important part in supporting A* (if more global considerations had supported some very different factor B, we might have had intuitional grounds for rejecting it). And that intuitional support for A could come from psychological evidence.

My argument, then, is entirely compatible with acknowledging top-down epistemic pressures, though I suspect that in many ethical domains, including discussion of trolley problems, they are in fact fairly weak.

The objection can therefore be defused. If it is to have real force against the argument, it must involve the claim that it's *simply impossible* to disentangle the epistemic contribution made by particular intuitions from that made by these other factors. But I doubt that this is a practicable (or even coherent) view of moral justification, and it seems to me to bear little resemblance to actual moral practice.

The step from (3) and (4) to (5) is a simple one. To reject (4) one would need to show that although our intuitions do often track general principles, that they do so is not any kind of *empirical* fact, indeed not something on which empirical evidence could even bear. I am not sure how one could defend such a claim, which comes dangerously close to suggesting that the empirical research actually being done in this area is incoherent.[31]

The armchair search for local coherence isn't simply a matter of assessing relations of coherence between a given set of intuitions about particular cases and our more general moral beliefs. The factors to which our intuitions are responding are typically opaque to introspection—our intuitions tell us that pushing is wrong, but not *why*.[32] And even in highly simplified and artificial examples such as Footbridge, there are numerous overlapping factors in play to which our intuitions might be responding. Philosophers form *conjectures* about these potential factors, conjectures which they then test, not only by assessing their coherence with particular intuitions and general beliefs, but also (and especially) by considering further cases. And these conjectures commit philosophers to what are in essence

[30] In some discussion of reflective equilibrium, 'intuition' is used to refer to our particular moral judgments. It thus makes perfect sense to speak of the iterative *revision* of intuition in light of principles and other epistemic factors. This is not what I mean by 'intuition', or what it typically means in current usage. On this usage it would be more accurate to say that top-down pressures give us reason to *reject* some intuitions, not to *revise* them. I'm grateful here to Simon Rippon.

[31] Perhaps it could be defended by appeal to worries about rule-following (see Kripke 1982). I don't have space here to address this large and unresolved issue. It suffices to say that if the rule-following considerations support scepticism about meaning and normativity, or radical particularism, then they undermine not just my argument, but also current ethical practice. Or it might be claimed that we possess direct perceptual openness to the moral properties around us: our intuitions track systematic patterns because there are such patterns in moral reality, *not* because of anything internal to our psychology. To the extent that this is meant to go beyond the truism that intuitive beliefs are non-inferential, I don't see how to make empirical sense of it.

[32] See Cushman et al. (2006) for empirical evidence that we often don't have introspective access to the factors which our intuitions track.

empirical predictions about how our intuitions would respond to the presence and absence of the conjectured factors in these further cases.

Moral philosophers certainly *are* happy to engage in empirical speculation about the psychological factors that underlie our intuitions once they have concluded (or strongly suspect) that our intuitions in the relevant domain are not reliable—once they have given up (1). Judith Thomson now speculates (in ways that echo Greene's views) that the intuitive difference between Sidetrack and Footbridge is due to differences in the directness of the harm caused.[33] This is clearly a testable empirical hypothesis, a matter for psychology to determine. But this would be a testable empirical hypothesis *whether or not* we reject (1). What our intuitions track, and whether they track a morally relevant property, are simply different questions.

4 Psychological evidence

All the premises of my argument, including (4), are essentially present in the passage from Kamm I quoted earlier.[34] And (5) simply follows from (3) and (4). Why doesn't Kamm take this further step?

We could accept (5) but think it uninteresting if we also held that

> Local coherence can identify the principles that our intuitions track *better* than empirical methods, or at least *just as accurately* and *effectively*.

This concedes that in ethical theorizing we in part seek to discover certain psychological facts, yet claims that we can best discover these facts from the armchair. But although we should expect that armchair reflection and psychological inquiry will often significantly overlap—if they didn't, strong sceptical consequences would follow—this suggestion is nevertheless wildly implausible.

If we think of ethical inquiry as a communal rather than solitary practice, where the intuitions that are taken on board are widely shared, stable and persistent, and constrained by common standards of epistemic screening, then ethical inquiry is likely to single out underlying patterns of covariance reasonably well, at least when

[33] Thomson, *ibid.* Notice also that when ethicists write as if the empirical accounts proposed by psychologists are inconsistent with their armchair views (e.g. Kamm 1998), they are implicitly accepting (4).

[34] Kamm also suggests another understanding of her methodology, on which "certain concepts that people have always worked with (even consciously) commit them—without their having realized it, consciously or unconsciously—to other concepts. The responses to cases reveal that one set of concepts and principles commit us to others, and these other concepts and principles can then be added on to the description of the underlying structure of the responses, but the structure was not always psychologically real." (Kamm 1993, 6) These remarks are not easy to interpret. Does Kamm have in mind the concept of 'wrongness', or perhaps of 'killing'? It's not likely that the content of such concepts determines answers to substantive moral questions. Or perhaps Kamm means that previous applications of these concepts to particular cases commit us to applying them in certain ways in new cases? (It's unlikely that Kamm thinks that our previous applications of moral concepts *logically* commit us to certain moral answers in these new cases—for then we should be able to identify the principle *without* having to consider new cases.) In any case, this passage doesn't reject (4)—it doesn't deny that our intuitions are generated by 'psychologically real' rules, only that these rules needed to be there 'all along'.

these are very robust. Viewed in this way, this part of armchair inquiry can be thought of as a kind of *qualitative* empirical research.

But this is a severely limited way of uncovering underlying causality (if it isn't, then many psychologists are wasting their time!). Consider first that since the notion of coherence employed in ethics is rather loose, two or more competing principles might cohere (or appear to cohere) equally well with our intuitions. This might even be the case at the end of ideal inquiry, after *all* possible scenario-intuition pairs have been considered. It's most certainly the case with respect to actual ethical inquiry, where in very many domains the method of local coherence has so far failed to identify a single principle that is agreed to account for our intuitions—it has failed to identify such a principle after over 40 years of reflection on the trolley problem!

Thus even if best local coherence was necessary for justification, it's clearly not sufficient. This means that at the very minimum, empirical evidence could serve as a *tie-breaker*, tipping the balance by identifying which of the competing moral principles in some domain is the one that is in fact causally operative.[35]

This however is too weak. Local coherence is merely a rough guide to correlations between patterns of non-moral properties and moral intuitions, and it is extremely implausible that it will on its own effectively and accurately identify the principles that are causally operative in the difficult cases that most interest ethicists. We should expect enough cases where the empirical evidence would identify one principle as causally operative, when the armchair search for local coherence identifies another, or, for that matter, fails to come up with any clear candidate.

To see this, consider first how empirical evidence might place general constraints on armchair inquiry. To start with, there are constraints of psychological plausibility. Some armchair principles might be simply *too complex* to plausibly underlie our intuitions.[36] But of course our understanding of psychology in general, and of moral psychology in particular, doesn't stop at such bland generalities. In deciding whether some suggested principle underlies our intuitions, we should make use of the best current psychological theories. We can appeal, for example, to the best available psychological accounts of the mechanisms that underlie moral intuitions. If evidence emerges that moral intuitions are generated by fairly simple processes, this would directly reduce the plausibility of proposed principles that are highly complex or demanding. Given that our understanding of moral psychology is still in its infancy, it might be that the prospects of some moral principles (and of the theories built around them) depend on the fortunes of certain empirical research

[35] Nichols (2005) compares ethical theory to an 'external' approach to linguistics, which is interested only in identifying the principles which are extensionally adequate to our intuitions—and there can be more than one—and isn't interested in *which* is psychologically realized. This is potentially misleading. Ethical theory isn't, of course, interested in *how* a principle is psychologically realized, but it *is* interested in identifying which one is *actually* being tracked, hence in which is causally operative (where that doesn't imply that the principle is *internally* represented).

[36] Some of principles defended by Kamm seem vulnerable to this worry—though Kamm could reply that the computations implicated in vision and language comprehension *are* highly intricate. Notice that this constraint doesn't show that highly elaborate moral principles couldn't be true, or even justifiable. What it might show is that, given human psychology, such principles cannot be justified by appeal to particular intuitions.

programmes. (It's not surprising then that Kamm has been most enthusiastic about Hauser's work, and most scathing about Greene's!)

But there are of course more direct means of testing empirical hypotheses about underlying causality. Psychologists can conduct experiments that directly test such hypotheses by studying the relevant patterns of intuitions in ordinary folk. Psychological evidence collected in this way would directly compete with the results of armchair reflection.[37] And it's doubtful that armchair coherence can distinguish as accurately as the powerful statistical methods of psychology between genuine and merely accidental patterns, or that the standard criteria of epistemic screening do more than weed out the mostly blatantly spurious intuitions.[38]

Moreover, ethical theorising is typically concerned with local coherence between principles and a small number of examples, sometimes only one of each type. It might thus be easily distorted by irrelevant accidental factors that are present in some examples. The more rigorous controls used in psychology, where many variants of the same example-type are used, better control for these and similar distorting influences. The same goes for the rather loose and inefficient ways that current ethical practice screens for individual idiosyncrasies in intuition. Although reported intuitions that are not shared by other ethicists are typically ignored, it's enough for some stray intuitions to be shared by several influential ethicists for them to dominate discussion for some time. Again the more rigorous methods of empirical research better control for such factors. Then there is the point already hinted above that local coherence simply can't distinguish genuine causality from mere correlation.

Empirical evidence at this level might not merely constrain or supplant the armchair search for local coherence. It might even make it redundant. If we can use empirical means to identify the principle that our intuitions track, then we have (defeasible) reason to believe it *even* when it conflicts with the principle singled out by local armchair coherence.

There already is a psychological research programme studying intuitions about trolley cases. If my argument is sound, this programme might eventually supersede part of the parallel armchair inquiry. Indeed, if my argument is sound, then ethicists should *encourage* (rather than be apprehensive about) similar programmes in *other* contested domains of morality, e.g. distributive justice or responsibility. It's not as if ethics can afford to dismiss a source of evidence that could potentially advance ethical theorising in virtually all domains. Even if such empirical evidence won't resolve longstanding ethical debates, it is hard to believe that it will not advance them in *any* way.

Notice that the empirical inquiry recommended by the present argument differs from current empirical inquiry in accepting the assumption that our intuitions are

[37] For example, when Scanlon rejects the DDE as an explanation of our intuitions about a range of cases, and when he suggests that there is no common principle that underlies them (Scanlon 2008, 4), he can be plausibly interpreted as committing himself to testable empirical claims that go counter to some influential psychological theories.

[38] Moreover, empirical methods would allow us to give some epistemic weight to effects on intuitions that are statistically significant but relatively weak. Such effects are simply ignored in current ethical practice, which is only sensitive to very robust patterns.

broadly reliable. This places a constraint on empirical theorising, in that it gives *priority* to those causal hypotheses that have at least prima facie moral plausibility. There is the danger that left to their own devices, psychologists would stop the inquiry *too early*.[39] This distinctive constraint leaves an important role to the armchair in the generation of such causal hypotheses and examples to test them—but of course psychologists also generate their hypotheses in the armchair.

Conversely, however, it's also possible that in a range of ethical disputes the relevant underlying factor has *already* been identified by armchair methods, but the search for some morally relevant difference continues precisely *because* of the assumption of reliability, the difficulty of giving up strong intuitions, and the limitations of the method of local coherence. Empirical evidence could help us bring closure to such disputes.

Finally, there is evidence at the subpersonal level. We saw earlier that the very same mapping function might be realized by focused emotional responses, dedicated computational modules, or domain general processes, or any combination of the above. Since on my argument what matters for ethical theorizing are only higher-level psychological facts about what mapping function generates our intuitions, how these facts are subpersonally realized makes no difference. The scientific findings about the neural processes that underlie moral judgement that have received so much attention are, from this perspective, *largely irrelevant*. What matters is whether our intuitions about trolleys track the DDE or some other principle, not whether affective or cognitive areas of the brain are involved in implementing this function.[40]

Still, lower level facts might play a useful role by offering *indirect evidence* about these higher level facts. For instance, if the DDE underlies our intuitions about killing, then brain areas typically involved in ascribing intentions should be associated with such intuitions. This is an empirically testable hypothesis.[41]

There is a more radical possibility at least worth mentioning. We saw that if some mapping function is psychologically realized, this doesn't mean that the principle being tracked needs to be explicitly represented. But it's at least possible that the moral principles underlying our intuitions *are* explicitly represented—perhaps even written in 'mentalese'. If this is the case, then it might be possible to identify the principle underlying some set of intuitions by deciphering what is encoded in the Language of Thought. This would offer us direct empirical access to the principles underlying our intuitions, making both local armchair coherence and traditional psychological methods redundant.[42]

[39] See e.g. Kamm (1998).

[40] Such subpersonal evidence might still be relevant for arguments that seek to undermine the *reliability* of our intuitions, though this is dubious. See Berker (2009) and Kahane (2011).

[41] See Berker (2009, 328) for a similar point. Berker however also inaccurately remarks that neuroscience can identify only correlations, not causation. This is true only of neuroimaging. Manipulation of activity in the relevant brain areas using e.g., transcranial magnetic stimulation, or identification of double dissociations in patient populations, can provide strong evidence for causation.

[42] When there are grounds for supposing that some pattern of intuitions is innate, evolutionary considerations could also provide indirect evidence about the factors that these intuitions track (see Singer 2005).

5 Clarifications

5.1 Not deriving an 'ought' from an 'is'

The argument I outlined ends with the epistemic conclusion that empirical evidence can support moral principles. If *this* argument is correct, then space is opened for a range of arguments for *substantive* conclusions. *Assuming* the conclusion of my main argument, such substantive arguments would start from sets of particular intuitions which, conjoined with empirical evidence about the properties these intuitions are tracking, would support some principle citing these properties:

(a) We have (undefeated) intuitions about Sidetrack, Footbridge and relevant cases.
(b) There is strong empirical evidence that these intuitions systematically respond to the difference between intending and merely foreseeing harm.

Therefore

(c) We have defeasible reason to believe the DDE.

This is *not* an argument from 'is' to 'ought', from facts about our psychology to moral conclusions. The argument *doesn't* claim that moral principles are true or justified simply because they reflect our innate biological nature, or our psychological make-up. Rather, it's an argument from *particular* oughts to a *general* ought. The role of empirical evidence in such arguments is to identify the common *grounds* of these particular oughts, something to which we obviously do not have introspective access. So our gaze remains firmly fixed on 'moral reality'; what the psychological evidence does is help us see *what it is we are seeing*. It has no independent epistemic (or moral) authority.[43]

5.2 Not a rejection of consistency

My argument might be misunderstood in another way. It in no way implies that we should ignore consistency in moral belief, or adopt principles that are not consistent with our intuitions.

We are asking what general principles to adopt in light of our intuitions. In many cases, we don't start with both general moral principles and specific intuitions, which we then mutually adjust until equilibrium is reached. Rather the principles we consider—principles such as the DDE—were proposed in the first case *in order* to account for some range of intuitions. The search for local coherence offers one way of deciding between such competing principles—of deciding both what principle to adopt *and* which intuitions to preserve or to reject. My argument suggests that empirical methods are another, and better, way of identifying the principles that

[43] It's thus inaccurate to say, as Daniels does, that when we aim to identify the principles that underlie some pattern of intuitions, our aim is merely descriptive (Daniels 1980). Indeed, to the extent that our initial intuitions are a priori, and that principles identified by empirical means also need to be scrutinised for their intrinsic plausibility, then experimental ethics as outlined here might even be compatible with a broadly *rationalist* moral epistemology (compare: the way computers are used in applied mathematics).

really underlie our intuitions. One potential result of such identification is that a different set of intuitions will be preserved. But the resulting set of intuitions and principles *will* strongly cohere. And, again, we might have reason to reject an empirically-identified principle if it does not cohere with our *other* moral principles and values, just as we might reject principles identified in the armchair on the same grounds.

5.3 Doesn't presuppose moral realism

My argument sets out from the assumption that our intuitions are reliable, that they track general moral patterns rather than serve as the building blocks out of which principles are constituted. This might suggest that the argument presupposes moral realism. Antirealists need not worry, and can return to their armchairs.

But in setting out the argument, I have just adverted to assumptions that I take to widely shape substantive ethical reflection. Different metaethical views would offer competing accounts of these assumptions. Since different aspects of our actual moral practice must presumably decide between these competing accounts, it's possible that only a realist view can fully account for these assumptions. If so, this is a problem for antirealist views, not for my argument.

The argument itself, however, doesn't presuppose moral realism. It requires only that there be empirical facts about what principles (if any) our intuitions are tracking. It says *nothing* about the semantics and metaphysics of moral propositions. If some response-dependent view is correct, then what our intuitions track are our psychological states (or our psychology in some ideal condition). And although on noncognitivist views moral propositions express rather than refer to such states, their (quasi-)truth would still be *correlated* with some complex psychological states. (If some antirealist view is true, it is *easier*, not harder, to see how our psychology could be a guide to what is morally true or correct.)

The relevant psychological states would presumably be our general conative attitudes in favour or against various types of acts. That we have such general attitudes is obvious enough, and it wouldn't be surprising if our intuitions about particular cases reliably (but imperfectly) reflected these more general attitudes. On many antirealist views, our interest in generating intuitions about new cases, and in epistemic screening, can be interpreted as means of uncovering these more general attitudes, and distinguishing them from merely accidental surface responses.[44]

5.4 Considered judgment and expert intuition

When empirical evidence is claimed to undermine some common philosophical appeals to intuitions, for example by showing them to be more diverse than assumed, or culturally specific, philosophers often object that these findings are

[44] That antirealist views can recognize a gap between intuitions about particular cases and more general moral truths is compatible with holding that there is a constitutive link between having some general conative attitude and our attitudes about various relevant particular things. Such a constitutive link would allow for many exceptions, and might only hold in ideal circumstances. Indeed, on some of these views our intuitions wouldn't themselves be conative attitudes.

irrelevant because (i) the intuitions philosophers appeal to are *considered* or *reflective* judgments, not the immediate responses studied by psychologists, and (ii) because these intuitions are the refined intuitions of trained *philosophers*, not the untutored intuitions of lay people.[45]

A similar objection might be levelled at my argument. Psychologists study the immediate intuitions of non-philosophers. But ethical theory appeals to the considered judgment of ethicists, who are better trained at making subtle conceptual distinctions and resisting the influence of irrelevant factors.

This objection makes an empirical assumption: that the intuitions of ethicists are systematically different from those of lay people. At least with respect to the basic trolley cases, the evidence doesn't support this assumption—expert and lay intuitions are in agreement.[46] And even if the intuitions of expert ethicists were systematically different, this would not yet show that they are more reliable. As psychologists would rightly point out, even if ethicists are better at drawing some distinctions, their intuitions might also be unconsciously biased by theoretical assumptions and motivations. (There are further questions. Who counts as a relevant expert—should we distinguish assistant, associate and emeritus intuitions? And what should we do when the intuitions of experts clash?)

I do not need to resolve these questions here.[47] Even if the intuitions of ethicists were different, and more reliable, this would do nothing to block my argument. All it would mean is that this argument best applies to the psychology of ethicists, not to the typical subjects of psychological experiments.[48]

It would actually be *unfortunate* if this was the case. It would mean that what we need to identify is the mapping function underlying *ethicists'* responses to trolley dilemmas. This would not be any less of an empirical question, to be studied by empirical means. But it would be a question that is far harder to study (or to get the psychologists to study). The consequence might be that we would still need to use the trusted old armchair to answer ethical questions, but this should be cause for regret, not celebration.

6 The scope of the argument

My argument establishes a foothold for experimental ethics. The scope of this empirical encroachment would be constrained by the extent to which ethical practice relies on the search for local coherence, as it is only this search that will be supplanted by empirical inquiry.

[45] See e.g. Williamson (2011).

[46] Cushman et al. (2006).

[47] Though see Schwizgebel and Cushman (forthcoming) for some evidence that the trolley intuitions of ethicists are not superior to those of lay people.

[48] Even this is too strong. The objection essentially denies that the assumption of the reliability of moral intuition applies to lay intuitions. It would be far more plausible to claim that ethicists' intuitions are more reliable, and therefore have *greater* evidential force, which is compatible with taking lay intuitions to still be of significant epistemic interest.

I have repeatedly emphasized that ethical theorising doesn't end with local coherence. Local coherence only gives defeasible support to a principle, which is then assessed for intrinsic plausibility, coherence with other principles and values, and its impact on other theoretical virtues, such as simplicity and economy. If we replace local coherence with empirical evidence, these global aspects of ethical reflection will remain intact—and can only be pursued in the armchair.

My argument applies to *any* piece of ethical theorising that, in at least *some* stage, appeals to local coherence to provide epistemic support to moral principles. It's clear enough that this methodological practice is widespread—it's by no means restricted to Frances Kamm. It dominates discussion of the trolley problem, and many similar discussions.

The exact spread of this practice is itself an empirical question. Our answer to this question would determine the horizontal reach of the argument: how *widely* it applies. The extent to which ethicists give weight to considerations that go beyond local coherence is a *separate* question which would determine the argument's vertical reach: what role it leaves for the armchair *when* it applies. Obviously there is far greater variance on this dimension, with Kamm representing one extreme. Although Kamm acknowledges that the principles identified through reflection on intuitions will eventually need to be made intelligible in light of general moral values, only little of her work is devoted to this second task. I actually suspect that she is not alone, and that much ethical theorising doesn't stray very far from questions about local coherence.

Now if there are ways of justifying moral principles without reference to particular intuitions, my argument leaves these unaffected. Perhaps the principle of utility (or that of beneficence) can be justified purely by reflection on its content—justified intuitively, but not by appeal to intuitions about particular cases. However, scrutiny of the intrinsic plausibility of the content of some general principle often plays a merely *negative* role. For example, it seems intrinsically *im*plausible that mere physical contact could make a moral difference—the distinction between intending and foreseeing doesn't ring the same alarm bells. Yet in such cases the epistemic role of reflection on principles is still entirely dependent on the *prior* positive support of particular intuitions. I very much doubt that anyone would endorse the DDE purely on the basis of reflection on the intrinsic plausibility of its content, if it was shown to get *no* support whatsoever from particular intuitions.

Finally, my argument is irrelevant to approaches to ethics that abstain from *any* appeal to particular intuitions—approaches that reject premise (1) without qualification. Consider however the two main ethical theories that have been claimed to have non-intuitional foundations, utilitarianism and Kantian ethics. Although these theories were first presented in forms that radically depart from some aspects of commonsense morality, they have over time gravitated closer and closer to common intuitions, their radical edge gradually ironed out. Some currently influential versions of these theories are explicitly defended by appeal to moral intuitions.[49] And even the few remaining radicals suffer, I believe, from inevitable lapses to their vow of abstinence: when uncompromising act utilitarians defend

[49] See e.g. Hooker's rule consequentialism (Hooker 2000) or Audi's intuitionist Kantianism (Audi 2001).

some conception of well-being, or address the non-identity problem, they can't help but consider intuitions about particular cases.

6.1 Not an armchair pyre

The idea that empirical research can challenge philosophical and ethical practice has been in the air for a while in naturalist circles. In its most common form, it has been expressed in sceptical arguments that attempt to show that our intuitions in some domain, or even in general, are deeply unreliable. Such arguments often challenge philosophers' common intuitions about some subject matter by showing, using surveys, that there is greater diversity in intuitions about it than philosophers have assumed.[50] A more controversial type of argument has tried to appeal to evidence from neuroimaging to debunk moral intuitions.[51]

The argument outlined in this paper is very different. Not only doesn't it challenge the reliability of our moral intuitions, it takes their basic reliability as its *premise*—it challenges current practice at one level up, the level of principles. But even here, its primary aim is positive: if correct, my argument suggests a better method for identifying moral principles than the armchair search for local coherence. Finally, my argument primarily appeals to empirical facts about functional causal claims, not to empirical facts about mere diversity in intuitions or about neural processes.

6.2 Morally irrelevant factors

It's worth noting, however, that some of these more negative arguments seem to implicitly presuppose the core of my argument. Greene and others, for example, have claimed that their experiments show that our intuitions about trolleys are really responsive to factors that are clearly morally irrelevant, such as whether some ways of killing involve intimate physical contact, and should be therefore dismissed. This is a sceptical argument. But set out carefully, it involves several steps, some of which overlap with the more positive argument developed here:

(A) Empirical evidence suggests that some set of moral intuitions is tracking property P.
(B) This casts doubt on the principles that ethicists have suggested to explain these moral intuitions, principles which do not cite P.
(C) P is morally irrelevant.

Therefore

(D) These moral intuitions are unreliable.

My basic argument is implicit in premises (A) and (B), which assume that armchair coherence without causality is defeated. But if we stopped here, the

[50] See Sinnott-Armstrong (2006), Knobe and Nichols (2008).

[51] Greene (2008), Singer (2005).

argument would be entirely compatible with taking our intuitions to be reliable. It at most challenges some principles identified in the armchair.

Yet those who develop such arguments, as well as those who try to rebut them, typically assimilate (B) and (C)—perhaps even assume that the empirical evidence casts doubt on armchair principles *because* it shows the operative factor to be morally irrelevant.[52] This is a mistake. The truth of (A) would be sufficient for (B). And premise (C) is not based on any empirical evidence. It's a product of the armchair, and itself based in intuition (at the general principle level).[53]

It is possible that empirical research of the kind I am recommending would reveal some moral intuitions to be tracking morally irrelevant properties. If this turned out to be common, we would have to conclude that our moral intuitions are unreliable. This is an open question. But we could also arrive at such a sceptical outcome from the armchair—(C) and (D) make *no* mention of our psychology. To repeat: we shouldn't confuse questions about the reliability of our intuitions with distinct questions about what principles underlie them.

6.3 Post hoc rationalization

It has sometimes been claimed, on the basis of empirical research, that the elaborate moral principles that ethicists investigate in the armchair are mere ex post rationalizations—even 'confabulations'.[54] The idea is that although our moral intuitions are actually tracking rather simple and morally irrelevant properties, ethicists engage in an elaborate form of self-deception and weave intricate but spurious theories in the attempt to make moral sense of these intuitions.

My argument suggests a better way to make this complaint against the ethicists. The complaint is simply that in theorizing about moral intuitions, ethicists have not taken constraints of psychological plausibility into account—that they have overlooked the empirical commitments of their methodology. We can avoid the offensive talk about self-deception and confabulation. And this version of the complaint leaves space for the possibility that the ethicists *are* right. It might be that our psychological machinery is more intricate than some psychologists currently assume.

7 Conclusion

Many years ago, John Rawls compared ethical theory to Chomskian linguistics. This analogy has recently been revived and investigated as an empirical hypothesis about the subpersonal underpinnings of morality.[55] But the analogy has another

[52] These two claims are often conflated with the distinct third claim that the subpersonal processes generating these intuitions are unreliable. This third claim belongs to a different kind of debunking argument that needn't appeal to moral premise (C). When the above argument works, however, then of course this further claim might *explain* why a morally irrelevant property is being tracked.

[53] See Berker (2009).

[54] See Greene (2008), and Kahneman's remarks on Kamm in Voorhoeve (2009).

[55] Mikhail (2007).

aspect which is of greater interest to ethics: it was also a claim about the nature of a key part of ethical theorising. Rawls suggested that when ethical theorists try to uncover the principles underlying their intuitions about particular cases, what they were doing is similar to what linguists do when they consider our grammatical intuitions and try uncover the syntactic structures that underlie them. This suggestion can be accepted independently of any further Chomskian baggage about 'moral grammar' or innateness.

There is a sense in which my argument echoes Rawls's analogy: there is a core component of ethical theorising that is concerned with what is ultimately an empirical question. This concern can be squared with the normative aims of ethics, given that for ethicists this empirical question about the properties tracked by our intuitions is driven by a presumption about the basic reliability of our moral intuitions. But linguistics, which in the sense meant by Rawls was *also* a kind of armchair inquiry, is a misleading model. This empirical question is best pursued not in the armchair, but in the field and at the laboratory, and the relevant discipline is not linguistics but psychology in the broad sense which also encompasses neuroscience and perhaps evolutionary psychology. And this is now not just an analogy, but a thriving actual empirical research programme.[56]

Discussing Rawls's analogy thirty years ago, Norman Daniels argued that, because the search for local coherence ultimately amounts to a descriptive question, it couldn't be central to ethics.[57] Now ethical theorising certainly doesn't end when we identify the principle that underlies some pattern of intuitions, and perhaps this is sometimes not even where it starts. But ethical theorising *very often* starts with the search for local coherence, and if my argument is correct then the psychological presuppositions of this armchair pursuit have far reaching implications for the practice of ethics.[58]

[56] It's worth signalling several ways in which the argument could be further extended. (1) To the extent that we can draw a competence/performance distinction in the moral domain, there might be empirical grounds to completely discount some set of *actual* intuitions and the principle they track and endorse instead the principle they *would've* tracked if we lacked some computational limitation. Discovering that certain intuitions are merely moral heuristics would have similar epistemic implications. (2) To the extent that some set of intuitions has its basis in a disposition selected by evolution, and we have strong evolutionary grounds for thinking that this disposition was originally selected in order to track some property X, this might give us grounds for endorsing X even if our intuitions in fact now seem to track some property Y. It might, however, be argued that when we trace the origins of intuitions to evolution we also undermine their justification, making questions about what properties these intuitions track irrelevant (see Kahane 2011).

[57] Daniels (1980, 25). See fn. 43.

[58] Can my argument be also deployed in other domains where intuitions play a role? In some domains, intuitions play no more than a heuristic role, and there are intuition-independent methods for justifying general beliefs. In others, it's perfectly clear what general principles the relevant intuitions are tracking, and empirical methods are redundant. The argument has greatest potential in domains where armchair appeal to particular intuitions plays a significant (or indispensable) yet inconclusive role in justifying general principles. Some other areas of philosophy may well meet this condition: if some account of knowledge is based on intuitions, and empirical evidence shows that these intuitions in fact track other factors, then surely this account of knowledge is challenged.

Acknowledgments I am grateful to several anonymous referees for useful comments. I have also greatly benefited from suggestions by audiences at Delft, Oxford and Amsterdam, and from written comments by Simon Rippon and Regina Rini, and, especially, by Selim Berker. Work on this paper was supported by a University Award from the Wellcome Trust (WT087208MF).

References

Audi, R. (2001). A Kantian intuitionism. *Mind, 110*(439), 601–635.
Berker, S. (2009). The normative insignificance of neuroscience. *Philosophy & Public Affairs, 37*(4), 293–329.
Cushman, F., Young, L., & Hauser, M. D. (2006). The role of conscious reasoning and intuition in moral judgments: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.
Dancy, J. (2006). *Ethics without principles*. Oxford: Oxford University Press.
Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy, 76*, 256–282.
Daniels, N. (1980). Some methods of ethics and linguistics. *Philosophical Studies, 37*, 21–36.
Dworkin, R. (1973). The original position. *The University of Chicago Law Review, 40*(3), 500–533.
Dwyer, S. (1999). Moral competence. In K. Murasugi & R. Stainton (Eds.), *Philosophy and linguistics*. Boulder, CO: Westview Press.
Greene, J. D. (2008). The secret joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral psychology: The neuroscience of morality* (pp. 35–79). Cambridge, MA: MIT Press.
Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.
Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400.
Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293*, 105–2108.
Hooker, B. (2000). *Ideal code*. Real World, Oxford: Oxford University Press.
Huemer, M. (2006). *Ethical intuitionism*. New York: Palgrave Macmillan.
Jackson, F., Pettit, P., & Smith, M. (2000). Ethical particularism and patterns. In B. Hooker & M. O. Little (Eds.), *Moral particularism*. Oxford: Oxford University Press.
Kahane, G. (2011). Evolutionary debunking arguments. *Noûs, 45*(1), 103–125.
Kamm, F. M. (1993). *Morality, mortality* (Vol. I). Oxford: Oxford University Press.
Kamm, F. M. (1998). Moral intuitions, cognitive psychology, and the harming-versus-not-aiding distinction. *Ethics, 108*, 463–488.
Kamm, F. M. (2006). *Intricate ethics*. Oxford: Oxford University Press.
Knobe, J., & Nichols, S. (Eds.). (2008). *Experimental philosophy*. New York: Oxford University Press.
Kripke, S. (1982). *Wittgenstein on rules and private language: An elementary exposition*. Cambridge, MA: Harvard University Press.
Mikhail, J. (2007). Universal moral grammar: theory, evidence, and the future. *Trends in Cognitive Science, 11*(4), 143–152.
Nichols, S. (2005). Innateness and moral psychology. In P. Carruthers, S. Laurence, & S. Stich (Eds.), *The innate mind: Structure and content*. New York: Oxford University Press.
Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.
Scanlon, T. (2008). *Moral dimensions: Permissibility, meaning, blame*. Cambridge, MA: Harvard Belknap.
Schwizgebel, E., & Cushman, F. (forthcoming). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind and Language*.
Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics, 9*, 331–352.

Sinnott-Armstrong, W. (2006). Moral intuitionism meets empirical psychology. In T. Horgan & M. Timmons (Eds.), *Metaethics after Moore*. Oxford: Oxford University Press.
Sosa, E. (2009). *Reflective knowledge: Apt belief and reflective knowledge* (Vol. II). Oxford: Oxford University Press.
Thomson, J. J. (1985). The trolley problem. *Yale Law Journal, 94*, 1395–1415.
Thomson, J. J. (2008). Turning the trolley. *Philosophy & Public Affairs, 36*, 359–374.
Voorhoeve, A. (2009). *Conversations on ethics*. Oxford: Oxford University Press.
Williamson, T. (2011). Philosophical expertise and the burden of proof. *Metaphilosophy, 42*, 215–229.